学校编码：10384　　分类号_____密级____
学号：23020081153217　　UDC____

厦门大学

硕　士　学　位　论　文

基于最大公共子图的中文Web文本分类研究

Research on Chinese Web Text Categorization Based on Maximum Common Subgraph

赖　兴　瑞

指导教师姓名：张东站　副教授
专　业　名　称：计算机软件与理论
论文提交日期：　2011　年　5　月
论文答辩时间：　2011　年　6　月
学位授予日期：　2011　年　　月

答辩委员会主席：_____
评　　阅　　人：_____

2011 年　　月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月

摘 要

随着网络信息技术的高速发展，Internet 上的 Web 页面数量呈指数增长，如何有效的组织和处理这些海量信息，如何更好地搜索、过滤和管理这些网络资源，成为一个亟待解决的问题。Web 文本挖掘技术就是解决上述问题的一种方法，它借鉴数据挖掘的基本思想和理论方法，从大量半结构化、异构的 Web 文档的集合中发现潜在的、有价值的知识。Web 文本分类是 Web 文本挖掘的重要技术，是一种快速、有效的组织网上海量信息的关键技术，是 Web 信息处理的基础，有着很高的研究价值和广泛的应用前景。

本文研究的对象是中文 Web 文本，目的是提高 Web 文本分类的精度和速度，主要针对中文 Web 文本的表示以及分类算法进行了深入地探讨。

Web 文档包含大量的与主题内容无关的噪音数据，因此本文提出了一种基于网页分块的主题信息自动提取算法。首先对 Web 文档依据布局标签分块构建文本内容块层次树，然后自底向上遍历层次树，计算每个块节点的语义属性和主题相关度，同时删除主题无关节点，最终通过遍历文本块层次树的最大内容节点路径，提取当前网页的主题信息。实验表明该主题信息提取算法对大多数中文门户网站的主题型网页均有效，适用性比较强。

传统的向量空间文本表示方法不能有效表示文本的结构信息，缺乏对文本特征词条上下文环境的考虑，因此本文探讨了 Web 文档的图表示方法、文档图之间距离度量选择等问题，并在此基础上发展了 KNN 算法，得到了基于最大公共子图的 Web 文本分类算法：MCS-KNN 算法。MCS-KNN 算法为每个 Web 文档生成表示图，通过计算两个 Web 文档表示图之间的相似度来计算两者的相似度，进而计算出待分类文档在训练集中的 K 近邻，根据 K 近邻的所属类别确定待分类文档的类别。实验表明，MCS-KNN 算法分类速度快，精度高，具有比 KNN 算法更优越的分类性能。

关键字：Web 文本分类；主题信息提取；最大公共子图

Abstract

With the rapid development of network information technology, the number of web pages on the Internet grows exponentially. It has become an urgen problem to be solved that how to organize and process these huge amounts of information effectively and how to search, filter and manage these network resources better. Web text mining is one of the solutions of these problems. By borrow ideas from the basic ideas and theoretical methods of data mining, it discovers the potential, valuable knowledge from large semi-structured, heterogeneous collection of web documents. Web text categorization is an important technology of text mining. It's a critical technology for organizing the mass online information, and it's the basis of web information processing. Web text categorization has high research value and broad application prospects.

This paper focused on Chinese web text, the purpose of which is to improve the accuracy and the speed of web text classification. In this paper, the representation of Chinese web text and its classification algorithm were discussed in depth.

For web document usually containing a large amount of noise that is irrelevant to its topic, this paper puts forward an algorithm for automatic extraction of topic information which is based on web partition. First, partition web document according to layout labels to build the hierarchical tree of text blocks. Second, traversing the hierarchical tree bottom-up, we figure out semantic attributes and theme correlativity of each block node, and remove irrelevant nodes. Finally, by traversing through the path of maximum content nodes of the hierarchical tree, we extract the topic information of current web page. Experiments show that this extraction algorithm is effective for topic pages of most Chinese web portals, while it has relatively strong applicability.

For that the traditional method of vector space text representation can not effectively capture the structure information of the text, and lack of consideration on

the context of feature terms, this paper first discussed problems in depth such as the graph representation model of web documents, how to measure the distance between two representation graph of two web documents. And then, we put forwards a web text categorization algorithm based on maximum common graph on the basis of the KNN algorithm, called MCS-KNN algorithm. The MCS-KNN algorithm first generates the representation graph for each web document. We calculate the similarity of two web documents by means of the similarity of their representation graphs. And then, figure out the *K* nearest neighbor of the unclassified web document among all the train instances, and decide the target class of current unclassified web document in line with categories of the *K* nearest neighbor. Experiments show that MCS-KNN algorithm has more superior classification performance than KNN algorithm for its fast speed and high precision.

Key Words: web text categorization; topic information extraction; maximum common subgraph

目　录

Contents

第一章 绪论

1.1 研究背景与意义

随着网络技术的高速发展，Internet 上网络信息资源呈指数级增长，据估计网页数量每 4 到 6 个月翻一翻，使得 Web 上的信息量以惊人的速度增长，Internet 包含了大量的信息资源，在这些大量的、异质的信息资源中，隐藏着具有巨大潜在价值的知识，所以它已经成为了世界性的图书馆，变成了各行各业人们交流思想、获取信息的平台。但是面对 Web 如此丰富的内容，巨大的数据量，加上万维网动态开放的特点，人们要去快速准确的寻找自己需要的信息不是一件容易的事情，通常需要耗费大量的人力和物力，人们面临着“信息爆炸”而“知识贫乏”的窘境。

由于网络上的信息大多以文本的形式表达，文本提供给用户大量丰富的信息，在这些信息中包含了潜在的、有巨大价值的知识，而面对如此庞大的文本资源，传统的文本分析处理工具已经无法满足现实需要的要求，人们迫切需要研究出有效的方法和手段从大规模文本信息资源中提取符合需要的简洁、精炼、可理解的知识。因此，Web 文本挖掘成为了数据挖掘中一个日益流行而重要的研究课题，是 Web 挖掘研究的重心。

Web 文本挖掘中最关键的技术是 Web 文本分类。网页分类是指在给定的分类体系下，根据网页文本的内容自动确定文本类别的过程，是一种典型的有指导学习的方法。Web 文本分类作为 Web 文本挖掘的重要技术和重要内容，有着越来越重要的意义，已经成为智能信息检索和处理领域的一个新兴和重要的研究方向。

现有的文本分类系统基本都是基于向量空间模型 VSM（Vector Space Model）。向量空间模型的基本思想是把文本表示成向量空间中的向量，采用向量之间的夹角余弦作为文本间的相似性度量。向量空间模型是一种不考虑特征项出现顺序的词袋文本表示模型，虽然带来了计算和操作上的便利，却损失了大量的文本结构

信息，缺乏对特征词条上下文环境的考虑，而这些文本结构信息或者上下文环境在自然语言中是至关重要的。向量空间模型的这些局限，客观上也限制了基于向量空间模型的传统文本分类算法分类性能的进一步提高。

与一般的数据结构相比，图能够表达更加丰富的语义。图结构能够模拟几乎所有事物间的联系，它能应用到半结构化和非结构化的数据挖掘中。建立 Web 文档的图表示模型，使之可以反映文档中特征词条、特征词条间的联系以及特征词条的共现程度等文本信息，并在此模型表示的基础上发展相应的文本分类方法是当前进一步提高文本分类性能的有益探索方向。

1.2 文本分类研究现状

1.2.1 国外研究现状

文本分类的研究可以追溯到上世纪六十年代，早期的文本分类主要是基于知识工程（Knowledge Engineering），通过手工定义一些规则来对文本进行分类，这种方法费时费力，且必须对某一领域有足够的了解，才能写出合适的规则。

到上世纪九十年代，随着网上在线文本的大量涌现和机器学习的兴起，大规模的文本（包括网页）分类和检索重新引起研究者的兴趣。文本分类系统首先通过在预先分类好的文本集上训练，建立一个判别规则或分类器，从而对未知类别的新样本进行自动归类。大量的结果表明它的分类精度比得上专家手工分类的结果，并且它的学习不需要专家干预，能适用于任何领域的学习，使得它成为目前文本分类的主流方法。

1971 年，Rocchio[1]提出了在用户查询中不断通过用户的反馈来修正类权重向量，来构成简单的线性分类器。1979 年，Van Rijsbergen[2]对信息检索领域的研究做了系统的总结，里面关于信息检索的一些概念，如向量空间模型（Vector Space Model）和评估标准如准确率(Precision)、召回率(Recall)，后来被陆续地引入文本分类中，文中还重点地讨论了信息检索的概率模型，而后来的文本分类研究大多数是建立在概率模型的基础上。

1992 年，Lewis 在他的博士论文[3]中系统地介绍了文本分类系统实现方法的